Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Multivariate Analysis 96 (2005) 219–236

Journal of Multivariate Analysis

www.elsevier.com/locate/jmva

On the consistency properties of linear and quadratic discriminant analyses

Santiago Velilla[a,*,1], Adolfo Hernández[b]

[a] *Departamento de Estadística, Universidad Carlos III de Madrid, 28903-Getafe, Madrid, Spain*
[b] *Department of Mathematical Sciences, University of Exeter, Exeter, Devon EX4 4QE, UK*

Received 26 March 2003
Available online 7 December 2004

Abstract

The limit behavior of the conditional probability of error of linear and quadratic discriminant analyses is studied under wide assumptions on the class conditional distributions. Results obtained may help to explain analytically the behavior in applications of linear and quadratic discrimination techniques.

AMS 1991 subject classification: 62H30; 62H99

Keywords: Bayes error; Conditional probability of misclassification; Consistent sample discriminant rules; Inverse regression models; Plug-in sample discriminant rules

1. Introduction

Consider a discriminant problem, where the goal is to assign an individual to one of a finite number of classes or groups $g_1, \ldots, g_k$ on the basis of p observed features $\mathbf{x} = (x_1, \ldots, x_p)'$. Let $\mathbf{D}_n = \{\mathbf{x}_{ij} : i = 1, \ldots, k, j = 1, \ldots, n_i\}$ be a training database of individuals previously classified, where $\mathbf{x}_{ij}$ is the jth individual in the ith class and $n = \sum_{i=1}^{k} n_i$ is the total sample size. Let also $\bar{\mathbf{x}}_i = \sum_{j=1}^{n_i} \mathbf{x}_{ij}/n_i$ and $\mathbf{S}_i = \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)$

* Corresponding author.
E-mail addresses: santiago.velilla@uc3m.es (S. Velilla), A.Hernandez@ex.ac.uk (A. Hernández).

[1] Research partially supported by *CICYT* Grant *BEC* 2002 – 03769 (Spain).

0047-259X/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2004.10.009

$(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)'/n_i$ be, respectively, the sample $p\times 1$ mean vector and $p\times p$ covariance matrix of the data in the ith group, $i=1,\ldots,k$. The *linear discriminant analysis* (LDA) rule assigns $\mathbf{x}$ to g_i when

$$(\mathbf{x}-\bar{\mathbf{x}}_i)'\mathbf{S}_p^{-1}(\mathbf{x}-\bar{\mathbf{x}}_i)=\min_{1\leqslant j\leqslant k}(\mathbf{x}-\bar{\mathbf{x}}_j)'\mathbf{S}_p^{-1}(\mathbf{x}-\bar{\mathbf{x}}_j)\,, \tag{1}$$

where $\mathbf{S}_p=\sum_{i=1}^k n_i\mathbf{S}_i/(n-k)=\sum_{i=1}^k\sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)'/(n-k)$ is the sample $p\times p$ *pooled* covariance matrix. A generalization of (1) is the *quadratic discriminant analysis* (QDA) rule

$$\begin{aligned}&\log(|\mathbf{S}_i|)+(\mathbf{x}-\bar{\mathbf{x}}_i)'\mathbf{S}_i^{-1}(\mathbf{x}-\bar{\mathbf{x}}_i)\\&\quad=\min_{1\leqslant j\leqslant k}[\log(|\mathbf{S}_j|)+(\mathbf{x}-\bar{\mathbf{x}}_j)'\mathbf{S}_j^{-1}(\mathbf{x}-\bar{\mathbf{x}}_j)]\,.\end{aligned} \tag{2}$$

For an introduction to LDA and QDA, see for example [17, Section 4.3].

Fisher [9] for the case of $k=2$ groups and Rao [28, Section 9c] for the general case $k>2$, gave an alternative derivation of LDA based on a concept of separation of populations. Let $\widehat{\mathbf{B}}=\sum_{i=1}^k n_i(\bar{\mathbf{x}}_i-\bar{\mathbf{x}})(\bar{\mathbf{x}}_i-\bar{\mathbf{x}})'/n$ and $\widehat{\mathbf{W}}=\sum_{i=1}^k\sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)'/n$ be, respectively, the sample $p\times p$ *between* groups and *within* groups dispersion matrices, where $\bar{\mathbf{x}}=\sum_{i=1}^k\sum_{j=1}^{n_i}\mathbf{x}_{ij}/n$ is the sample mean vector. The idea is to find the directions that maximize the Fisher–Rao discriminant criterion

$$\max_{\mathbf{a}\in\mathbb{R}^p}\frac{\mathbf{a}'\widehat{\mathbf{B}}\mathbf{a}}{\mathbf{a}'\widehat{\mathbf{W}}\mathbf{a}}\,, \tag{3}$$

and hence the spread of the $\bar{\mathbf{x}}_i$ relative to the within class variability. For $j=1,2,\ldots$, the *discriminant directions* $\widehat{\mathbf{a}}_j$ can be obtained recursively as the eigenvectors of $\widehat{\mathbf{W}}^{-1}\widehat{\mathbf{B}}$ (see e.g. [29, Chapter 5]). For $r\leqslant p$, define also the *canonical* or *discriminant coordinates* $\widehat{\mathbf{y}}_r=\widehat{\mathbf{A}}_r'(\mathbf{x}-\bar{\mathbf{x}})$ of the feature vector $\mathbf{x}$, where $\widehat{\mathbf{A}}_r=(\widehat{\mathbf{a}}_1,\widehat{\mathbf{a}}_2,\ldots,\widehat{\mathbf{a}}_r)$ is a $p\times r$ matrix. Assigning $\widehat{\mathbf{y}}_r$ to the closest centroid $\widehat{\mathbf{m}}_{r,i}=\widehat{\mathbf{A}}_r'(\bar{\mathbf{x}}_i-\bar{\mathbf{x}})$, $i=1,\ldots,k$, leads to the *reduced rank linear discriminant analysis* $(RLDA)$ rule

$$\|\widehat{\mathbf{y}}_r-\widehat{\mathbf{m}}_{r,i}\|^2=\min_{1\leqslant j\leqslant k}\|\widehat{\mathbf{y}}_r-\widehat{\mathbf{m}}_{r,j}\|^2\,, \tag{4}$$

where $\|.\|$ is the usual euclidean norm. For an overview of the construction and properties of RLDA, see [17, Section 4.3].

Despite the customary interpretation of LDA and QDA as sample maximum likelihood discriminant rules related to normal populations, Hastie et al. [17, p. 89] comment on the good behavior of the linear and quadratic rules in a diverse set of applications. There is not a clear explanation for this phenomenon, and the available answers depend on empirical experiments and heuristic arguments. For example, Hastie et al. [17, p. 89] suggest that the reason does not seem to lie in the approximate gaussianity of the class conditional densities, but rather on the fact that the data can only support simple linear or quadratic separation boundaries. Hastie and Zhu [18, p. 180] remark that criterion (3) depends only on location and dispersion, and this may provide a partial justification for the *robustness* of LDA under non-gaussian conditions. On the other hand, and as suggested by Johnson and Wichern [19, Chapter 11], not much is known about the behavior of RLDA in applications. In particular,

there is not a well established procedure for choosing in practice the number of canonical coordinates r (see e.g. [10, Section 7.3]).

The aim of this paper is to study the limit behavior of the conditional probability of misclassification of both linear and quadratic discriminant analyses. This behavior can give, when the sample size n is moderately large, some analytical answers for the issues related in the previous paragraph. Section 2 establishes notation, reviews previous literature, and gives some background and motivation for our results. Section 3 is devoted to LDA and QDA, and Section 4 to RLDA. Section 5 contains some final comments.

2. Background and motivation

This section introduces the notation that will be used throughout the paper. Since our results are motivated by adequate *pseudo plug-in* representations of LDA, QDA and RLDA, the consistency properties of these sample criteria as parametric plug-in rules are also reviewed.

2.1. Notation

Let $\mathbf{g}$ denote the class membership of the individual under study and, for $i = 1, \ldots, k$, put $\pi_i = P[\mathbf{g} = i] > 0$ for the ith class prior probability and $f_i(\mathbf{x})$ for the ith class conditional density. Any solution to the classification problem defined by the pairs $(\pi_i, f_i(\mathbf{x}))$, $i = 1, \ldots, k$, is given by a discriminant rule $r(\mathbf{x}) = \sum_{i=1}^{k} i\, I_{R_i}(\mathbf{x})$, where the subsets $R_1, \ldots, R_k$ form a partition of $\mathbb{R}^p$ and $I_{R_i}(.)$ is the indicator function of R_i. The notation implies that $\mathbf{x}$ is assigned to g_i when $r(\mathbf{x}) = i$ or, equivalently, when $\mathbf{x}$ belongs to R_i. Given $(\mathbf{x},\mathbf{g})$, rule $r(\mathbf{x})$ is in error when $r(\mathbf{x}) \neq \mathbf{g}$ and the probability of misclassification $L[r(\mathbf{x})] = P[r(\mathbf{x}) \neq \mathbf{g}]$ is

$$L[r(\mathbf{x})] = 1 - \sum_{i=1}^{k} P[\mathbf{g} = i] P[\mathbf{x} \in R_i \mid \mathbf{g} = i] = 1 - \sum_{i=1}^{k} \pi_i \int_{R_i} f_i(\mathbf{x})\, d\mathbf{x}\,. \tag{5}$$

The optimal or Bayes rule, that is, the rule that minimizes the functional $L[r(\mathbf{x})]$ of (5), is given by the partition $R_i^* = \{\mathbf{x} : \pi_i f_i(\mathbf{x}) = \max_{1 \leqslant j \leqslant k} \pi_j f_j(\mathbf{x})\}$, $i = 1, \ldots, k$ (see e.g. [29, Chapter 6]). According to (5), the probability of misclassification of $r^*(\mathbf{x}) = \sum_{i=1}^{k} i\, I_{R_i^*}(\mathbf{x})$ is the optimal or Bayes error

$$L^* = L[r^*(\mathbf{x})] = 1 - \sum_{i=1}^{k} \pi_i \int_{R_i^*} f_i(\mathbf{x})\, d\mathbf{x}\,. \tag{6}$$

If $P[\pi_i f_i(\mathbf{x}) = \pi_j f_j(\mathbf{x})] = 0$ for all $i \neq j$, the Bayes rule is unique except for sets of probability zero [1, Chapter 6; 24, Chapter 1]. That is, if $s^*(\mathbf{x}) = \sum_{i=1}^{k} i\, I_{S_i^*}(\mathbf{x})$ is also optimal, then $P[s^*(\mathbf{x}) = r^*(\mathbf{x})] = 1$ and the partitions R_i^* and S_i^* are *equivalent*, that is, $P[\mathbf{x} \in R_i^* \triangle S_i^*] = 0$, $i = 1, \ldots, k$, where $\triangle$ is the symmetric difference of subsets.

In general both π_i and $f_i(\mathbf{x})$ are unknown, so the rules used in practice are of the form $\widehat{r}_n(\mathbf{x}) = \sum_{i=1}^{k} i\, I_{\widehat{R}_{i,n}}(\mathbf{x})$, where the subsets $\widehat{R}_{i,n}$, $i = 1, \ldots, k$, depend on the training

database $\mathbf{D}_n$. The appropriate measure of error of a sample rule $\widehat{r}_n(\mathbf{x})$ is its conditional probability of misclassification $L_n = P[\widehat{r}_n(\mathbf{x}) \neq \mathbf{g} \mid \mathbf{D}_n]$. From (5) and (6), if the pair $(\mathbf{x}, \mathbf{g})$ is independent of $\mathbf{D}_n$,

$$L_n = 1 - \sum_{i=1}^{k} \pi_i \int_{\widehat{R}_{i,n}} f_i(\mathbf{x})\, d\mathbf{x}\,, \tag{7}$$

and then the random variable L_n satisfies $0 \leqslant L^* \leqslant L_n \leqslant 1$. Following Devroye, Györfi and Lugosi [8, Chapter 6], the sequence of sample rules $\{\widehat{r}_n(\mathbf{x})\}$ is (Bayes risk) weakly or strongly consistent when, as n goes to infinity, L_n converges in probability or almost everywhere (a.e.) to the optimum L^*.

2.2. Consistency of LDA, QDA and RLDA as parametric plug-in rules

A common procedure for constructing sample rules in practice is the *plug-in* approach. The idea is to consider a partition of the form $\widehat{R}_{i,n} = \{\mathbf{x} : \widehat{\pi}_i \widehat{f}_{i,n}(\mathbf{x}) = \max_{1 \leqslant j \leqslant k} \widehat{\pi}_j \widehat{f}_{j,n}(\mathbf{x})\}$, $i = 1, \ldots, k$, where $\widehat{\pi}_i$ and $\widehat{f}_{i,n}(\mathbf{x})$ are estimators computed from $\mathbf{D}_n$ of π_i and $f_i(\mathbf{x})$, respectively. The priors are usually estimated by a fixed set of probabilities, for example $\pi_i = 1/k$, $i = 1, \ldots, k$, or by the sample proportions $\widehat{\pi}_i = n_i/n \to \pi_i$ a.e. Following Glick [11, Section 5], if, for each $i = 1, \ldots, k$, $\widehat{f}_{i,n}(\mathbf{x})$ is a probability density that converges to $f_i(\mathbf{x})$ a.e. for almost all $\mathbf{x} \in \mathbb{R}^p$, then the associated sequence of plug-in rules is strongly consistent. Therefore [11, example 4], if the class conditional densities are specified by k parametric families $f_i(\mathbf{x}) = h_i(\mathbf{x}; \mathbf{\Lambda})$, $i = 1, \ldots, k$, $\widehat{\mathbf{\Lambda}}_n$ is an estimator of $\mathbf{\Lambda}$ such that $\widehat{\mathbf{\Lambda}}_n \to \mathbf{\Lambda}$ a.e., and each $h_i(\mathbf{x}; \mathbf{\Lambda})$ is a continuous function of $\mathbf{\Lambda}$, the sequence of rules determined by the estimated densities $\widehat{f}_{i,n}(\mathbf{x}) = h_i(\mathbf{x}; \widehat{\mathbf{\Lambda}}_n)$, $i = 1, \ldots, k$, is strongly consistent. A natural choice for $\widehat{\mathbf{\Lambda}}_n$ is the maximum likelihood (ML) estimator of $\mathbf{\Lambda}$.

As an application of the previous result, LDA is strongly consistent when the π_i are equal and the $f_i(\mathbf{x})$ are multivariate normal $N_p(\boldsymbol{\mu}_i, \mathbf{\Sigma})$ with a common positive definite (p.d.) covariance matrix $\mathbf{\Sigma}$, $i = 1, \ldots, k$. To see this, notice that (1) is equivalent to criterion $\widehat{f}_{i,n}(\mathbf{x})/k = \max_{1 \leqslant j \leqslant k} \widehat{f}_{j,n}(\mathbf{x})/k$, where $\widehat{f}_{i,n}(\mathbf{x}) = (2\pi)^{-p/2} \mid \widehat{\mathbf{W}} \mid^{-1/2} \exp[-(\mathbf{x} - \overline{\mathbf{x}}_i)' \widehat{\mathbf{W}}^{-1} (\mathbf{x} - \overline{\mathbf{x}}_i)/2]$, $i = 1, \ldots, k$, that in turn is a plug-in version of the optimal rule $f_i(\mathbf{x})/k = \max_{1 \leqslant j \leqslant k} f_j(\mathbf{x})/k$ (or equivalently $(\mathbf{x} - \boldsymbol{\mu}_i)' \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) = \min_{1 \leqslant j \leqslant k} (\mathbf{x} - \boldsymbol{\mu}_j)' \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_j)$) with parameters $\boldsymbol{\mu}_i$ and $\mathbf{\Sigma}$ replaced by their *ML* estimators $\overline{\mathbf{x}}_i$ and $\widehat{\mathbf{W}}$, respectively. Similarly, QDA is strongly consistent when the $f_i(\mathbf{x})$ are $N_p(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i)$ with different covariance matrices $\mathbf{\Sigma}_i$, $i = 1, \ldots, k$. This asymptotic optimality under multivariate normality of LDA and QDA, in the presence of homoscedasticity and heteroscedasticity respectively, is explained in [24, Chapter 3]. See also [11, example 4].

On the other hand, when the $f_i(\mathbf{x}) \sim N_p(\boldsymbol{\mu}_i, \mathbf{\Sigma})$, $i = 1, \ldots, k$, and the rank $r_0 = r(\mathbf{B}) \leqslant \min(k-1, p)$ of the population between groups dispersion matrix $\mathbf{B} = Var[E(\mathbf{x} \mid \mathbf{g})] = \sum_{i=1}^{k} (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})'/k$ is known, where $\boldsymbol{\mu} = E(\mathbf{x}) = \sum_{i=1}^{k} \boldsymbol{\mu}_i/k$, Hastie and Tibshirani [15, Section 3] obtain the ML estimators $\widehat{\boldsymbol{\mu}}_i(r_0)$ and $\widehat{\mathbf{W}}(r_0)$ of $\boldsymbol{\mu}_i$ and $\mathbf{\Sigma}$ respectively, computed under the restriction $r_0 = r(\mathbf{B})$. See also [3]. These estimators can be written in terms of the spectral decomposition

$$\widehat{\mathbf{W}}^{-1/2} \widehat{\mathbf{B}} \widehat{\mathbf{W}}^{-1/2} = \widehat{\mathbf{C}} \widehat{\mathbf{D}} \widehat{\mathbf{C}}\,, \tag{8}$$

where $\widehat{\mathbf{C}}$, with columns $\widehat{\gamma}_j$, is a $p \times p$ orthogonal matrix of eigenvectors, and $\widehat{\mathbf{D}}$ is a $p \times p$ diagonal matrix of eigenvalues $\widehat{\lambda}_1 \geqslant \widehat{\lambda}_2 \geqslant \cdots \geqslant \widehat{\lambda}_p \geqslant 0$. It can be checked that

$$[\mathbf{x} - \widehat{\boldsymbol{\mu}}_i(r_0)]' \widehat{\mathbf{W}}^{-1}(r_0)[\mathbf{x} - \widehat{\boldsymbol{\mu}}_i(r_0)] = \|\widehat{\mathbf{A}}'_{r_0}(\mathbf{x} - \overline{\mathbf{x}}_i)\|^2 + \widehat{Q}(\mathbf{x}) \,, \tag{9}$$

where, writing $\widehat{\mathbf{C}}_{r_0} = (\widehat{\gamma}_1, \widehat{\gamma}_2, \ldots, \widehat{\gamma}_{r_0})$, $\widehat{\mathbf{C}}_{(r_0)} = (\widehat{\gamma}_{r_0+1}, \ldots, \widehat{\gamma}_p)$ and $\widehat{\mathbf{D}}_{(r_0)} = diag(\widehat{\lambda}_{r_0+1}, \ldots, \widehat{\lambda}_p)$, $\widehat{\mathbf{A}}_{r_0} = \widehat{\mathbf{W}}^{-1/2}\widehat{\mathbf{C}}_{r_0} = (\widehat{\mathbf{a}}_1, \widehat{\mathbf{a}}_2, \ldots, \widehat{\mathbf{a}}_{r_0})$ is a $p \times r_0$ matrix of discriminant directions normalized by conditions $\widehat{\mathbf{a}}'_j \widehat{\mathbf{W}} \widehat{\mathbf{a}}_k = 1$ $(j = k)$ and $= 0$ $(j \neq k)$, $\|\widehat{\mathbf{A}}'_{r_0}(\mathbf{x} - \overline{\mathbf{x}}_i)\|^2 = \|\widehat{\mathbf{y}}_{r_0} - \widehat{\mathbf{m}}_{r_0,i}\|^2$ is as defined in (4), and $\widehat{Q}(\mathbf{x}) = \| [\mathbf{I}_{p-r_0} + \widehat{\mathbf{D}}_{(r_0)}]^{-1/2} \widehat{\mathbf{C}}'_{(r_0)} \widehat{\mathbf{W}}^{-1/2}(\mathbf{x} - \overline{\mathbf{x}}) \|^2$ is a quadratic term that depends on $\mathbf{x}$ and the training sample, but not on the class index i.

Consequently, under the assumption $r_0 = r(\mathbf{B})$, the rule of (4) corresponding to r_0 canonical coordinates is equivalent to the *ML* plug-in rule defined by the estimated normal densities $\widehat{f}_{i,n}(\mathbf{x}) = (2\pi)^{-p/2} \mid \widehat{\mathbf{W}}(r_0) \mid^{-1/2} \exp\{-[\mathbf{x} - \widehat{\boldsymbol{\mu}}_i(r_0)]' \widehat{\mathbf{W}}^{-1}(r_0)[\mathbf{x} - \widehat{\boldsymbol{\mu}}_i(r_0)]/2\}$, $i = 1, \ldots, k$. Then, RLDA is consistent for $r = r_0$. In practice, this result is of limited interest, since, with the exception of a two group problem in which $r_0 = 1$, the rank of the matrix $\mathbf{B}$ is in general unknown.

2.3. *Pseudo plug-in representations of LDA, QDA and RLDA*

Consider now a continuous and strictly decreasing function $h_0(.)$ from [0,∞) to [0,∞) which is such that $h_0(t) > 0$ for all $t \geqslant 0$. For convenience, suppose also that $h_0(\mathbf{u}'\mathbf{u})$ is a probability density in $\mathbf{u} \in \mathbb{R}^p$. Since $\widehat{\mathbf{W}} = (n-k)\mathbf{S}_p/n$, the LDA rule in (1) is equivalent to

$$\widehat{h}_{i,n}(\mathbf{x})/k = \max_{1 \leqslant j \leqslant k} \widehat{h}_{j,n}(\mathbf{x})/k \,, \tag{10}$$

where $\widehat{h}_{i,n}(\mathbf{x}) = \mid \widehat{\mathbf{W}} \mid^{-1/2} h_0[(\mathbf{x} - \overline{\mathbf{x}}_i)' \widehat{\mathbf{W}}^{-1}(\mathbf{x} - \overline{\mathbf{x}}_i)]$, $i = 1, \ldots, k$. Observe that this *sample equivalence* holds regardless of the model assumed for both the class prior probabilities and the class conditional densities. Notice also that, since $h_0(.)$ is arbitrary, the random function $\widehat{h}_{i,n}(\mathbf{x})$ above is not necessarily a consistent estimator of $f_i(\mathbf{x})$, $i = 1, \ldots, k$. Then, criterion (10) can be appropriately viewed as a *pseudo plug-in* representation of LDA. Similarly, when $\log |\mathbf{S}_i| \cong c$, $i = 1, \ldots, k$, where c is some fixed constant, the QDA rule in (2) is approximately equivalent to the pseudo plug-in classification criterion

$$\widehat{h}_{i,n}(\mathbf{x})/k = \max_{1 \leqslant j \leqslant k} \widehat{h}_{j,n}(\mathbf{x})/k \,, \tag{11}$$

where now $\widehat{h}_{i,n}(\mathbf{x}) = |\mathbf{S}_i|^{-1/2} h_0[(\mathbf{x} - \overline{\mathbf{x}}_i)' \mathbf{S}_i^{-1}(\mathbf{x} - \overline{\mathbf{x}}_i)]$, $i = 1, \ldots, k$.

In the same fashion, for r canonical coordinates, the RLDA rule of (4) can be represented as

$$\widehat{h}_{i,n}(r; \mathbf{x})/k = \max_{1 \leqslant j \leqslant k} \widehat{h}_{j,n}(r; \mathbf{x})/k \,, \tag{12}$$

where the sample density $\widehat{h}_{i,n}(r; \mathbf{x}) = |\widehat{\mathbf{W}}|^{-1/2} h_0[\widehat{Q}_i(r; \mathbf{x})]$ is determined by the quadratic

$$\widehat{Q}_i(r; \mathbf{x}) = \|\widehat{\mathbf{y}}_r - \widehat{\mathbf{m}}_{r,i}\|^2 + \|\widehat{\mathbf{C}}'_{(r)} \widehat{\mathbf{W}}^{-1/2}(\mathbf{x} - \overline{\mathbf{x}})\|^2, \tag{13}$$

$i = 1, \ldots, k$, where $\widehat{\mathbf{C}}_{(r)} = (\widehat{\gamma}_{r+1}, \widehat{\gamma}_2, \ldots, \widehat{\gamma}_p)$ is as defined in decomposition (8). The function $\widehat{Q}_i(r; \mathbf{x})$ of (13) is similar in structure to the decomposition (9) above, obtained in the normal case by Hastie and Tibshirani [15] when $r = r_0 = r(\mathbf{B})$.

In what follows, we investigate, under general assumptions for the class conditional densities $f_i(\mathbf{x})$, $i = 1, \ldots, k$, the limit behavior of the conditional probabilities of misclassification of the pseudo plug-in rules (10)–(12). This approach is new and different from the consistency properties of LDA, QDA and RLDA reviewed in the previous subsection, where an specific parametric normal model was supposed for the $f_i(\mathbf{x})$, $i = 1, \ldots, k$. The results obtained can offer some insight into the robustness properties of LDA and QDA (Section 3), as well as offering some guidelines for choosing in practice the number of directions in RLDA (Section 4).

3. Asymptotic properties of LDA and QDA

Observe first, in connection with representation (10) for LDA, that, if no particular assumption is made on the form of the class conditional densities, $\bar{\mathbf{x}}_i$ and $\widehat{\mathbf{W}}$ are not in general ML estimators of the population class mean $\boldsymbol{\mu}_i = E(\mathbf{x} \mid \mathbf{g} = i)$ and population within dispersion matrix $\mathbf{W}$, respectively. However, if the training sample $\mathbf{D}_n = \{\mathbf{x}_{ij} : i = 1, \ldots, k, j = 1, \ldots, n_i\}$ is formed by i.i.d. observations and the feature vector $\mathbf{x} = (x_1, \ldots, x_p)'$ satisfies $E(x_j^2) < +\infty$, $j = 1, \ldots, p$, using the law of the large numbers it follows that, as $n \to \infty$, $\bar{\mathbf{x}}_i \to \boldsymbol{\mu}_i$ a.e., $i = 1, \ldots, k$, and $\widehat{\mathbf{W}} \to \mathbf{W}$ a.e. Consequently, by continuity of $h_0(.)$,

$$\widehat{h}_{i,n}(\mathbf{x}) \to h_i(\mathbf{x}) = |\mathbf{W}|^{-1/2} h_0[(\mathbf{x} - \boldsymbol{\mu}_i)'\mathbf{W}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)], \tag{14}$$

a.e. for almost all $\mathbf{x}$, where $h_i(\mathbf{x})$ is an elliptically symmetric density that is not necessarily equal to $f_i(\mathbf{x})$, $i = 1, \ldots, k$. Similarly, the random functions used in representation (11) for QDA are such that $\widehat{h}_{i,n}(\mathbf{x}) = |\mathbf{S}_i|^{-1/2} h_0[(\mathbf{x} - \bar{\mathbf{x}}_i)'\mathbf{S}_i^{-1}(\mathbf{x} - \bar{\mathbf{x}}_i)] \to h_i(\mathbf{x}) = |\mathbf{V}_i|^{-1/2} h_0[(\mathbf{x} - \boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)]$, where $\mathbf{V}_i = Var(\mathbf{x} \mid \mathbf{g} = i)$ is the ith p.d. class conditional dispersion matrix, $i = 1, \ldots, k$.

The asymptotic properties of rules (10) and (11) are then a consequence of the following result, that characterizes the limit behavior of the conditional probability of error $L_n = P[\widehat{r}_n(\mathbf{x}) \neq \mathbf{g} \mid \mathbf{D}_n]$ of a sequence of sample rules $\{\widehat{r}_n(\mathbf{x})\}$ determined by a partition of the form

$$\widehat{R}_{i,n} = \{\mathbf{x} : \widehat{\pi}_i \widehat{h}_{i,n}(\mathbf{x}) = \max_{1 \leqslant j \leqslant k} \widehat{\pi}_j \widehat{h}_{j,n}(\mathbf{x})\}, \tag{15}$$

$i = 1, \ldots, k$, where the $\widehat{\pi}_i \to \pi_i$ a.e., and the $\Omega \times \mathbb{R}^p$ measurable random functions $\{\widehat{h}_{i,n}(\mathbf{x})\}$, where Ω is the underlying probability space, are such $\widehat{h}_{i,n}(\mathbf{x}) \to h_i(\mathbf{x})$ a.e. for almost all $\mathbf{x} \in \mathbb{R}^p$, where $h_i(\mathbf{x})$ is an arbitrary density function, $i = 1, \ldots, k$.

Proposition 3.1. *Under the previous assumptions on the sequences* $\{\widehat{\pi}_i\}$ *and* $\{\widehat{h}_{i,n}(\mathbf{x})\}$, $i = 1, \ldots, k$, *if* $P[\pi_i h_i(\mathbf{x}) = \pi_j h_j(\mathbf{x})] = 0$ *for all* $i \neq j$, *then, as* $n \to \infty$,

$$L_n \to 1 - \sum_{i=1}^{k} \pi_i \int_{R_i(\mathbf{h})} f_i(\mathbf{x})\, d\mathbf{x} \quad a.e., \tag{16}$$

where $R_i(\mathbf{h}) = \{\mathbf{x} : \pi_i h_i(\mathbf{x}) = \max_{1 \leqslant j \leqslant k} \pi_j h_j(\mathbf{x})\}$, $i = 1, \ldots, k$.

Proposition 3.1 can be easily established applying to definition (7) of L_n a well-known result on interchanging limits and integrals of random functions due to Glick [12] (see also Lemma 3.1.3 in [27, p. 191]), and the proof is therefore omitted. Although clearly related to the results of Glick [11] mentioned in Section 2.2 on the consistency of plug-in rules, this proposition is different in nature, since, as in representations (10) and (11) above, the $\widehat{h}_{i,n}(\mathbf{x})$ are not necessarily consistent estimators of the class conditional densities $f_i(\mathbf{x})$, $i = 1, \ldots, k$. Observe also that, in general, the right-hand side of expression (16) will be larger than the Bayes error L^*.

As a corollary, by comparing (16) to expression (6) for the optimal error, and using uniqueness of the Bayes rule when $P[\pi_i f_i(\mathbf{x}) = \pi_j f_j(\mathbf{x})] = 0$ for all $i \neq j$, the sequence of pseudo plug-in rules $\{\widehat{r}_n(\mathbf{x})\}$ considered in Proposition 3.1 will be strongly consistent if and only if $R_i^* = \{\mathbf{x} : \pi_i f_i(\mathbf{x}) = \max_{1 \leqslant j \leqslant k} \pi_j f_j(\mathbf{x})\} = R_i(\mathbf{h})$, $i = 1, \ldots, k$, that is, if and only if the limit and optimal partitions coincide. Clearly, a sufficient condition for this to occur is $h_i(\mathbf{x}) = f_i(\mathbf{x})$, $i = 1, \ldots, k$. This is however not necessary, since situations exist in which the partitions are identical but the densities $h_i(\mathbf{x})$ and $f_i(\mathbf{x})$ are not (see e.g. Section 3.2 below). Phrased differently, the consistency of a sequence of pseudo plug-in rules depends on the structure of the optimal partition, rather than on the specific probability model assumed for the pairs $(\pi_i, f_i(\mathbf{x}))$, $i = 1, \ldots, k$.

In general, the limit densities $h_i(\mathbf{x})$ will not be known. However, there may exist a complete characterization of the partition $R_i(\mathbf{h}) = \{\mathbf{x} : \pi_i h_i(\mathbf{x}) = \max_{1 \leqslant j \leqslant k} \pi_j h_j(\mathbf{x})\}$, $i = 1,\ldots, k$. This allows to identify conditions for the class conditional densities under which $R_i^* = \{\mathbf{x} : \pi_i f_i(\mathbf{x}) = \max_{1 \leqslant j \leqslant k} \pi_j f_j(\mathbf{x})\} = R_i(\mathbf{h})$, $i = 1,\ldots, k$, and thus the sequence of sample rules (15) is asymptotically optimal. This could be interesting in practice, for example when trying to decide on the use of a given sample rule that admits a pseudo plug-in representation of the form (15).

Applications of these principles in connection with LDA and QDA are studied next. For the rest of this paper, the class priors are assumed to be equal, that is, in all the results above $\widehat{\pi}_i = \pi_i = 1/k$, $i = 1, \ldots, k$.

3.1. LDA

As seen in (14), the random functions $\widehat{h}_{i,n}(\mathbf{x})$ used in representation (10) for LDA converge, if second order moments exist, to $h_i(\mathbf{x}) = |\ \mathbf{W}\ |^{-1/2}\ h_0[(\mathbf{x} - \boldsymbol{\mu}_i)'\mathbf{W}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)]$, $i = 1, \ldots, k$. Therefore, by Proposition 3.1, the conditional probability of error of LDA

converges a.e. to

$$1-\frac{1}{k}\sum_{i=1}^{k}\int_{R_i(\mathbf{h})} f_i(\mathbf{x})\,d\mathbf{x}\,, \tag{17}$$

where, from the strict monotonicity of $h_0(.)$, the subset $R_i(\mathbf{h}) = \{\mathbf{x} : h_i(\mathbf{x})/k = \max_{1\leqslant j\leqslant k} h_j(\mathbf{x})/k\}$ coincides with $\{\mathbf{x} : (\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{W}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) = \min_{1\leqslant j\leqslant k}(\mathbf{x}-\boldsymbol{\mu}_j)'\mathbf{W}^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\}$, $i = 1,\ldots,k$. Consequently, comparing (6) and (17), LDA will be consistent if and only if the optimal partition R_i^* is determined minimizing the Mahalanobis distances of the feature vector $\mathbf{x}$ to the population class means $\boldsymbol{\mu}_i$, $i = 1,\ldots,k$, in the metric defined by the $p\times p$ matrix $\mathbf{W} = E[Var(\mathbf{x}\mid\mathbf{g})] = \sum_{i=1}^{k}\mathbf{V}_i/k$.

As an application of this result, suppose that the class conditional densities are of the form

$$f_i(\mathbf{x}) = |\boldsymbol{\Sigma}|^{-1/2} f_0[(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)]\,, \tag{18}$$

$i = 1,\ldots,k$, where $\boldsymbol{\Sigma}$ is a $p\times p$ p.d. matrix, and $f_0(.)$ is an strictly decreasing function in (a subset of) $[0,\infty)$, not necessarily continuous, and such that $f_0(\mathbf{u}'\mathbf{u})$ is a density function in $\mathbf{u}\in\mathbb{R}^p$. Possible (continuous) choices for $f_0(.)$ are the multivariate normal, $f_0(t) = (2\pi)^{-p/2}\exp(-t/2)$, mixtures of normals with the same dispersion shape, $f_0(t) = (2\pi)^{-p/2}[(1-\varepsilon)\exp(-t/2)+\varepsilon\sigma^{-p}\exp(-t/2\sigma^2)]$, $0<\varepsilon<1$, $\sigma>0$, the multivariate Student's t_m distribution with $m>2$ degrees of freedom, $f_0(t) = c(m,p)[1+(t/m)]^{-(m+p)/2}$, where $c(m,p)>0$ is some adequate constant, and the general multivariate Pearson Type II or Type VII distributions considered by Cooper [6]. A useful description of (18) is given by the *inverse location regression* model [5, p. 158]

$$\mathbf{x}\mid\mathbf{g}\overset{D}{=} i = \boldsymbol{\mu}_i + \boldsymbol{\Sigma}^{1/2}\mathbf{u}\,, \tag{19}$$

$i = 1,\ldots,k$, where $\mathbf{u}$ is a random vector independent of the class label $\mathbf{g}$ with density $f_0(\mathbf{u}'\mathbf{u})$.

If second-order moments exist, $E(\mathbf{x}\mid\mathbf{g}=i) = \boldsymbol{\mu}_i$ and $\mathbf{V}_i = Var(\mathbf{x}\mid\mathbf{g}=i) = a\boldsymbol{\Sigma}$, $i = 1,\ldots,k$, where $a>0$ is a constant independent of the class index i [26, p. 34]. Therefore, $\mathbf{W} = E[Var(\mathbf{x}\mid\mathbf{g})] = a\boldsymbol{\Sigma}$, and then, under (18), the optimal partition R_i^* is determined minimizing $(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$ or, equivalently, $(\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{W}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$, $i = 1,\ldots,k$. This fact was noted by Glick [13, Section 4], generalizing a previous result of Day [7] (see also [24, p. 238]).

Proposition 3.1 allows then to extend immediately to the *whole* family (18) the strong consistency under normality of LDA, obtained in Section 2.2 as a consequence of a result by Glick [11, example 4] on the consistency of parametric plug-in rules. A partial converse is also true, since, by (17), LDA will be consistent *only* when the optimal partition is determined minimizing the Mahalanobis distances $(\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{W}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$, $i = 1,\ldots,k$. We believe that this is a new aspect of LDA, which indicates that, in some sense, (18) is the *largest* class under which rule (1) should be considered for its use in practice.

3.2. Remarks and comparison with previous work

Observe first that, although leading to the same partition, in general the limit densities $h_i(\mathbf{x}) = |\ \mathbf{W}\ |^{-1/2}\ h_0[(\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{W}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)]$ in (14) will be different from the class conditional densities $f_i(\mathbf{x})$ in (18), $i = 1,\ldots,k$.

Notice also that the type of limit properties that can be obtained for LDA under (18), exploiting Proposition 3.1 and the identity $R_i^* = R_i(\mathbf{h}), i = 1,\ldots,k$, are wider and different than the ones that would follow by applying to the estimated model densities

$$\widehat{f}_{i,n}(\mathbf{x}) = |\ \widehat{\mathbf{W}}/a\ |^{-1/2}\ f_0[(\mathbf{x}-\overline{\mathbf{x}}_i)'(\widehat{\mathbf{W}}/a)^{-1}(\mathbf{x}-\overline{\mathbf{x}}_i)]\,, \tag{20}$$

$i = 1,\ldots,k$, the parametric consistency result of Glick [11, example 4] mentioned in Section 2.2. This is because, although criterion (1) is clearly equivalent to the plug-in rule $\widehat{f}_{i,n}(\mathbf{x})/k = \max_{1\leqslant j\leqslant k}\widehat{f}_{j,n}(\mathbf{x})/k$ defined by the random functions of (20), and $\overline{\mathbf{x}}_i \to \boldsymbol{\mu}_i$ and $\widehat{\mathbf{W}}/a \to \mathbf{W}/a = \boldsymbol{\Sigma}$, the estimator $\widehat{f}_{i,n}(\mathbf{x})$ does not necessarily converge to the class conditional density $f_i(\mathbf{x}) = |\boldsymbol{\Sigma}|^{-1/2}\ f_0[(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)]$ in (18), since the measurable function $f_0(.)$ is not supposed to be continuous. Recall also that the constant $a = a(f_0) > 0$ will be in general unknown.

3.3. QDA

To proceed with the analysis, suppose that the p.d. class covariance matrices $\mathbf{V}_i = Var(\mathbf{x} \mid \mathbf{g} = i)$ are such that $\log|\mathbf{V}_i| \cong c$, $i = 1,\ldots,k$, where c is some fixed constant. This condition is quite flexible, since even if the determinants $|\mathbf{V}_i|$ are large and different, they will tend to be closer in the log scale. As $\log|\mathbf{S}_i| \to \log|\mathbf{V}_i| \cong c$, $i = 1,\ldots,k$, the approximate representation (11) for QDA in terms of the random functions $\widehat{h}_{i,n}(\mathbf{x}) = |\ \mathbf{S}_i\ |^{-1/2}\ h_0[(\mathbf{x}-\overline{\mathbf{x}}_i)'\mathbf{S}_i^{-1}(\mathbf{x}-\overline{\mathbf{x}}_i)]$, $i = 1,\ldots,k$, follows.

As seen earlier, $\widehat{h}_{i,n}(\mathbf{x}) \to h_i(\mathbf{x}) = |\ \mathbf{V}_i\ |^{-1/2}\ h_0[(\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)]$, $i = 1,\ldots,k$, so, by Proposition 3.1, the conditional probability of error of the pseudo plug-in rule (11) converges a.e. as $n \to \infty$ to

$$1 - \frac{1}{k}\sum_{i=1}^{k}\int_{R_i(\mathbf{h})} f_i(\mathbf{x})\,d\mathbf{x}\,,$$

where now, using condition $\log|\mathbf{V}_i| \cong c$, $i = 1,\ldots,k$, the subset $R_i(\mathbf{h}) = \{\mathbf{x} : h_i(\mathbf{x})/k = \max_{1\leqslant j\leqslant k} h_j(\mathbf{x})/k\}$ is approximately determined by minimizing the Mahalanobis distances $(\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$, $i = 1,\ldots,k$.

As an application, suppose that the class conditional densities are of the form

$$f_i(\mathbf{x}) = |\boldsymbol{\Sigma}_i|^{-1/2} f_0[(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)]\,, \tag{21}$$

$i = 1,\ldots,k$, where the $\boldsymbol{\Sigma}_i$ are a collection of $p\times p$ p.d. matrices, and $f_0(.)$ is as defined in (18). This family includes as a particular case the distributions considered by Cooper [6]. A convenient representation of (21) is given by the *inverse location-scale regression* model [5, p. 160]

$$\mathbf{x} \mid \mathbf{g} \overset{D}{=} i = \boldsymbol{\mu}_i + \boldsymbol{\Sigma}_i^{1/2}\mathbf{u}\,, \tag{22}$$

$i = 1, \ldots, k$, where the error term $\mathbf{u}$ is as in (19). If moments of second order exist, $E(\mathbf{x} \mid \mathbf{g} = i) = \boldsymbol{\mu}_i$ and $\mathbf{V}_i = Var(\mathbf{x} \mid \mathbf{g} = i) = a\boldsymbol{\Sigma}_i$, $a > 0$. Therefore, noticing that $\log|\boldsymbol{\Sigma}_i| = \log|\mathbf{V}_i| - p\log(a) \cong c - p\log(a)$, $i = 1,\ldots, k$, under (21) the optimal partition R_i^* is approximately determined by minimizing $(\mathbf{x} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)$ or, equivalently, $(\mathbf{x} - \boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)$, $i = 1, \ldots, k$.

By Proposition 3.1, QDA will be then (approximately) consistent not only under normality, but also under the whole class (21). A partial converse also follows, since for QDA to be consistent, the optimal partition must be found (approximately) by minimizing the Mahalanobis distances $(\mathbf{x} - \boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)$, $i = 1, \ldots, k$. A general family for the $f_i(\mathbf{x})$ under which this property holds is given by (21).

3.4. Applications

The results obtained in Sections 3.1 and 3.3 can be used to justify analytically some aspects of the behavior in applications of LDA and QDA. McLachlan [24, Section 5.6.1] reports conclusions from simulation studies for continuous feature data. For sample sizes n large enough, the linear and quadratic rules seem to work well when the class conditional densities are (elliptically) *symmetric* but not necessarily gaussian. However, both rules are affected when, given the model for the $f_i(\mathbf{x})$, the Bayes rule is not determined by minimizing the Mahalanobis distances, $(\mathbf{x} - \boldsymbol{\mu}_i)'\mathbf{W}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)$ or $(\mathbf{x} - \boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)$, $i = 1,\ldots, k$, respectively, as for example when a skewed lognormal distribution is used to generate the class conditional densities. This type of empirical findings are a direct consequence of our previous results on the connection existing between consistency of LDA and QDA and the associated form of the optimal partition.

As seen above, a natural framework for considering the use of LDA and QDA in practice is given by the inverse regression models (19) and (22), respectively. These models have an structure that depends on general characteristics of the multivariate class conditional distributions $\mathbf{x} \mid \mathbf{g} = i, i = 1, \ldots, k$, such as: *location* ($\boldsymbol{\mu}_i$), *dispersion*, either homoscedastic ($\boldsymbol{\Sigma}$) or heteroscedastic ($\boldsymbol{\Sigma}_i$), and a density $f_0(\mathbf{u}'\mathbf{u})$ for the random vector $\mathbf{u}$, that is decreasing and spherically *symmetric*, but not necessarily known, continuous or multivariate normal.

Consequently, rather than standard parametric plug-in rules associated to an estimated *continuous* probability model, LDA and QDA can be considered, in some sense, as *nonparametric* rules associated to the elementary inverse regression models (19) and (22), respectively, in which (elliptical) symmetry is a key factor. We believe that this is a new aspect of the linear and quadratic rules. In particular, the property for LDA formalizes the comments by Hastie and Zhu [18, p. 180], mentioned in the introduction, on the importance of location and dispersion for a correct performance in applications of the linear rule.

Finally, models (19) and (22) could be used as a *first* approximation for the class conditional distributions whenever their characteristics seem adequate for the classification problem at hand. As an illustration, consider the *wave form* data. This is an artificial classification problem with $k = 3$ groups and $p = 21$ variables, introduced by Breiman et al. [2, p. 49] and studied extensively in the literature of discriminant analysis. Let $h_1(i) = \max(6 - |i - 11|, 0)$, $h_2(i) = h_1(i - 4)$ and $h_3(i) = h_1(i + 4)$ be the shifted wave form functions and, for $j = 1, 2, 3$, define the 21×1 vector $\mathbf{h}_j = (h_j(i) : 1 \leqslant i \leqslant 21)$.

The probabilistic structure of $\mathbf{x}$ is

$$\mathbf{x} \mid \mathbf{g} = i \overset{D}{=} \boldsymbol{\mu}_i(\varepsilon) + \mathbf{u}\,, \tag{23}$$

$i = 1, \ldots, 3$, where $\varepsilon \sim U(0,1)$ is independent of $\mathbf{u} \sim N_{21}(\mathbf{0}, \mathbf{I}_{21})$ and $\boldsymbol{\mu}_i(\varepsilon) = \varepsilon \mathbf{h}_j + (1-\varepsilon)\mathbf{h}_k$, where $(j,k) = (1,2)$ for $i = 1$, $(j,k) = (1,3)$ for $i = 2$, and $(j,k) = (2,3)$ for $i = 3$. All the priors are set to $\pi_i = 1/3$. Model (23) can be approximated *to first order* by the inverse location $\mathbf{x} \mid \mathbf{g} = i = \boldsymbol{\mu}_i + \mathbf{u}$, where $\boldsymbol{\mu}_i = E[\boldsymbol{\mu}_i(\varepsilon)] = \boldsymbol{\mu}_i(1/2) = (\mathbf{h}_j + \mathbf{h}_k)/2$, $i = 1$, 2, 3. This approximation amounts to replacing the random vector $\boldsymbol{\mu}_i(\varepsilon) = \varepsilon \mathbf{h}_j + (1-\varepsilon)\mathbf{h}_k$ in (23), where $0 \leqslant \varepsilon \leqslant 1$, by the midpoint of the segment that connects $\mathbf{h}_j$ with $\mathbf{h}_k$. LDA and QDA are consistent under the latter approximate model and should be then expected to behave correctly in this problem, as least as preliminary classification methods. In fact, Hastie et al. [16] found that, in this data set, LDA outperformed their nonlinear methods in terms of misclassification error.

4. Asymptotic properties of RLDA

This section analyzes, as a function of the number r of canonical coordinates considered, the asymptotic behavior of the conditional probability of misclassification of the reduced linear rule (4). Motivated by the consistency results obtained for LDA in the previous section, an inverse location model of the form (18)–(19) is assumed for the class conditional densities $f_i(\mathbf{x})$, $i = 1, \ldots, k$. As before, the random vector $\mathbf{u}$ will have finite second-order moments.

4.1. Preliminaries

Recalling the notation of (8) for the spectral decomposition $\widehat{\mathbf{W}}^{-1/2}\widehat{\mathbf{B}}\widehat{\mathbf{W}}^{-1/2} = \widehat{\mathbf{C}}\widehat{\mathbf{D}}\widehat{\mathbf{C}}$, the jth eigenvalue $\widehat{\lambda}_j = \widehat{\boldsymbol{\gamma}}_j'\widehat{\mathbf{W}}^{-1/2}\widehat{\mathbf{B}}\widehat{\mathbf{W}}^{-1/2}\widehat{\boldsymbol{\gamma}}_j = \sum_{i=1}^k n_i[\widehat{\boldsymbol{\gamma}}_j'\widehat{\mathbf{W}}^{-1/2}(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})]^2/n$ calibrates the separation existing between the standardized class centroids $\widehat{\mathbf{W}}^{-1/2}(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})$, $i = 1, \ldots, k$, after projecting onto $\widehat{\boldsymbol{\gamma}}_j = \widehat{\mathbf{W}}^{1/2}\widehat{\mathbf{a}}_j$, $j = 1,\ldots, p$, where $\widehat{\mathbf{a}}_j$ is the jth discriminant direction. For each $1 \leqslant j \leqslant p$, we will partition $\widehat{\mathbf{C}} = (\widehat{\mathbf{C}}_j \mid \widehat{\mathbf{C}}_{(j)})$, where $\widehat{\mathbf{C}}_j = (\widehat{\boldsymbol{\gamma}}_1, \widehat{\boldsymbol{\gamma}}_2,\ldots, \widehat{\boldsymbol{\gamma}}_j)$ is of $p \times j$ and $\widehat{\mathbf{C}}_{(j)} = (\widehat{\boldsymbol{\gamma}}_{j+1},\ldots, \widehat{\boldsymbol{\gamma}}_p)$ is of $p \times (p-j)$.

Put now $q = \min(k-1, p)$. If $r_0 = r(\mathbf{B}) \leqslant q$ is the (unknown) rank of the matrix $\mathbf{B} = \sum_{i=1}^k (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})'/k$, the spectral decomposition

$$\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2} = \mathbf{C}\mathbf{D}\mathbf{C}'\,, \tag{24}$$

the population counterpart of (8), is given by the $p \times p$ orthogonal matrix of eigenvectors $\mathbf{C} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2,\ldots, \boldsymbol{\gamma}_p)$ and the $p \times p$ diagonal matrix of eigenvalues $\mathbf{D} = diag(\lambda_1,\ldots, \lambda_{r_0}, 0,\ldots, 0)$, where $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_{r_0} > 0$. For convenience, we will assume that all the non-null eigenvalues are simple, that is, $\lambda_j > \lambda_{j+1}$ for $1 \leqslant j \leqslant r_0$. As before, we will partition $\mathbf{C} = (\mathbf{C}_j \mid \mathbf{C}_{(j)})$, where $\mathbf{C}_j = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2,\ldots, \boldsymbol{\gamma}_j)$ is of $p \times j$ and $\mathbf{C}_{(j)} = (\boldsymbol{\gamma}_{j+1},\ldots, \boldsymbol{\gamma}_p)$ is of $p \times (p-j)$.

In the continuous case, the ith sample dispersion matrix $\mathbf{S}_i$ will be p.d. with probability one for all $n_i \geqslant p+1$, $i = 1, \ldots, k$, and the same is true for $\widehat{\mathbf{W}} = \sum_{i=1}^k n_i \mathbf{S}_i/n$. Also, the rank of the sample $p \times p$ between groups dispersion matrix $\widehat{\mathbf{B}} = \sum_{i=1}^k n_i(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})'/n$

can be seen to be equal to q with probability one. This property holds independently of the true value of $r_0 = r(\mathbf{B})$.

The eigenvalues of (8) satisfy then $\widehat{\lambda}_1 \geqslant \widehat{\lambda}_2 \geqslant \cdots \geqslant \widehat{\lambda}_q > 0 = \widehat{\lambda}_{q+1} = \cdots = \widehat{\lambda}_p$. This inequality implies $\widehat{\mathbf{C}}'_{(q)}\widehat{\mathbf{W}}^{-1/2}\overline{\mathbf{x}}_i = \widehat{\mathbf{C}}'_{(q)}\widehat{\mathbf{W}}^{-1/2}\overline{\mathbf{x}}$, $i = 1, \ldots, k$. As a consequence, following for example the derivation in [24, pp. 91 and 186–187], RLDA coincides with LDA for all $r \geqslant q$. This is a well-known result that is mentioned here mainly to emphasize that, while there are exactly $r_0 = r(\mathbf{B})$ canonical coordinates in the population, in the continuous case the possible number of sample canonical coordinates is $1 \leqslant r \leqslant q$. In fact, when the number of groups k is relatively large as compared to p, it may well occur that $r_0 = r(\mathbf{B}) \ll r(\widehat{\mathbf{B}}) = q$.

4.2. Convergence results

As seen in expression (12) of Section 2.3, RLDA is, for $1 \leqslant r \leqslant q$ canonical coordinates, equivalent to a pseudo plug-in criterion defined in terms of the sample densities

$$\widehat{h}_{i,n}(r; \mathbf{x}) = |\widehat{\mathbf{W}}|^{-1/2} h_0[\widehat{Q}_i(r; \mathbf{x})]\,,$$

$i = 1, \ldots, k$, where the quadratic $\widehat{Q}_i(r; \mathbf{x}) = \|\widehat{\mathbf{y}}_r - \widehat{\mathbf{m}}_{r,i}\|^2 + \|\widehat{\mathbf{C}}'_{(r)}\widehat{\mathbf{W}}^{-1/2}(\mathbf{x} - \overline{\mathbf{x}})\|^2$ is given in (13). Observe first that, since $\|\widehat{\mathbf{y}}_r - \widehat{\mathbf{m}}_{r,i}\|^2 = \|\widehat{\mathbf{C}}'_r\widehat{\mathbf{W}}^{-1/2}(\mathbf{x} - \overline{\mathbf{x}}_i)\|^2 = (\mathbf{x} - \overline{\mathbf{x}}_i)'\widehat{\mathbf{W}}^{-1/2}\widehat{\mathbf{C}}_r\widehat{\mathbf{C}}'_r\widehat{\mathbf{W}}^{-1/2}(\mathbf{x} - \overline{\mathbf{x}}_i)$ and $\widehat{\mathbf{C}}_{(r)}\widehat{\mathbf{C}}'_{(r)} = \mathbf{I}_p - \widehat{\mathbf{C}}_r\widehat{\mathbf{C}}'_r$, the limit behavior of the random functions $\widehat{h}_{i,n}(r; \mathbf{x})$ above depends on the $p \times p$ matrix $\widehat{\mathbf{C}}_r\widehat{\mathbf{C}}'_r$, the orthogonal projection operator onto the column space of $\widehat{\mathbf{C}}_r = (\widehat{\gamma}_1, \widehat{\gamma}_2, \ldots, \widehat{\gamma}_r)$.

If second-order moments exist, $\widehat{\mathbf{W}}^{-1/2}\widehat{\mathbf{B}}\widehat{\mathbf{W}}^{-1/2} \to \mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$ a.e., so using well-known results (see e.g. [31, Lemma 2.1, p. 726]), it follows that, when $\lambda_r > \lambda_{r+1}$,

$$\widehat{\mathbf{C}}_r\widehat{\mathbf{C}}'_r \to \mathbf{C}_r\mathbf{C}'_r \quad \text{a.e.}, \tag{25}$$

where the matrix $\mathbf{C}_r = (\gamma_1, \gamma_2, \ldots, \gamma_r)$ is as defined in the population spectral decomposition (24). Similarly, $\widehat{\mathbf{C}}_{(r)}\widehat{\mathbf{C}}'_{(r)} = \mathbf{I}_p - \widehat{\mathbf{C}}_r\widehat{\mathbf{C}}'_r \to \mathbf{I}_p - \mathbf{C}_r\mathbf{C}'_r = \mathbf{C}_{(r)}\mathbf{C}'_{(r)}$. Notice also that, since $r_0 = r(\mathbf{B})$, condition $\lambda_r > \lambda_{r+1}$ can only be satisfied for $1 \leqslant r \leqslant r_0$.

4.2.1. *Case* $1 \leqslant r \leqslant r_0$

From (25), $\widehat{Q}_i(r; \mathbf{x}) \to Q_i(r; \mathbf{x}) = \|\mathbf{C}'_r\mathbf{W}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}_i)\|^2 + \|\mathbf{C}'_{(r)}\mathbf{W}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})\|^2$, $i = 1, \ldots, k$. Thus,

$$\widehat{h}_{i,n}(r; \mathbf{x}) = |\widehat{\mathbf{W}}|^{-1/2} h_0[\widehat{Q}_i(r; \mathbf{x})] \to h_i(r; \mathbf{x}) = |\mathbf{W}|^{-1/2} h_0[Q_i(r; \mathbf{x})], \tag{26}$$

a.e. for almost all $\mathbf{x} \in \mathbb{R}^p$, $i = 1, \ldots, k$.

Consequently, by Proposition 3.1, the conditional probability of error $L_n(r)$ of RLDA converges a.e. as $n \to \infty$ to

$$L(r) = 1 - \frac{1}{k}\sum_{i=1}^{k} \int_{R_i(\mathbf{h}_r)} f_i(\mathbf{x})\, d\mathbf{x}\,, \tag{27}$$

where, using (26), the subsets $R_i(\mathbf{h}_r) = \{\mathbf{x} : h_i(r; \mathbf{x})/k = \max_{1 \leqslant j \leqslant k} h_j(r; \mathbf{x})/k\}$ are determined minimizing the population counterparts $\|\mathbf{C}'_r\mathbf{W}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}_i)\|^2$ of the sample

quantities $\|\widehat{\mathbf{y}}_r - \widehat{\mathbf{m}}_{r,i}\|^2 = \|\widehat{\mathbf{C}}_r'\widehat{\mathbf{W}}^{-1/2}(\mathbf{x} - \overline{\mathbf{x}}_i)\|^2$ considered in (4), $i = 1, \ldots, k$. On the other hand, the probabilities of error $L(r)$ of (27), $1 \leqslant r \leqslant r_0$, satisfy

$$L(1) > L(2) > \cdots > L(r_0) = L^* . \tag{28}$$

The proof of (28) follows similar lines as in the normal case and is therefore omitted (see e.g. [24, Section 3.9]).

Therefore, as a conclusion from (27) and (28), $L_n(r_0)$ converges a.e. to L^*. That is, for $r = r_0$ sample canonical coordinates, RLDA is asymptotically optimal under the whole family (18). This result extends the consistency under normality of the reduced linear rule for $r = r_0$, mentioned in Section 2.2 in connection with the ML decomposition (9).

4.2.2. Case $r_0 < r \leqslant q$

If $r_0 < q$ and the number r of sample canonical coordinates used exceeds r_0, the quadratic $\widehat{Q}_i(r; \mathbf{x}) = \|\widehat{\mathbf{y}}_r - \widehat{\mathbf{m}}_{r,i}\|^2 + \|\widehat{\mathbf{C}}_{(r)}'\widehat{\mathbf{W}}^{-1/2}(\mathbf{x} - \overline{\mathbf{x}})\|^2$ can be decomposed additively in the form $\widehat{Q}_i(r; \mathbf{x}) = \widehat{Q}_i(r_0; \mathbf{x}) + \widehat{D}_i(r_0, r; \mathbf{x})$, where, for $r_0 < r \leqslant q$,

$$\widehat{D}_i(r_0, r; \mathbf{x}) = \sum_{j=r_0+1}^{r} [\widehat{\boldsymbol{\gamma}}_j'\widehat{\mathbf{W}}^{-1/2}(\mathbf{x} - \overline{\mathbf{x}}_i)]^2 - \sum_{j=r_0+1}^{r} [\widehat{\boldsymbol{\gamma}}_j'\widehat{\mathbf{W}}^{-1/2}(\mathbf{x} - \overline{\mathbf{x}})]^2 , \tag{29}$$

$i = 1, \ldots, k$. Using the elementary bound $[\widehat{\boldsymbol{\gamma}}_j'\widehat{\mathbf{W}}^{-1/2}(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})]^2 \leqslant (n_i/n)^{-1}\widehat{\lambda}_j \to k\lambda_j = 0$ $(j > r_0)$, and the Cauchy–Schwartz inequality, the difference $\widehat{D}_i(r_0, r; \mathbf{x})$ in (29) can be seen to converge to zero a.e. for almost all $\mathbf{x}$, $i = 1, \ldots, k$.

Therefore, for $r_0 < r \leqslant q$, the random function $\widehat{h}_{i,n}(r; \mathbf{x}) = |\widehat{\mathbf{W}}|^{-1/2}h_0[\widehat{Q}_i(r; \mathbf{x})]$ of (12) converges to $h_i(r_0; \mathbf{x}) = |\mathbf{W}|^{-1/2}h_0[Q_i(r_0; \mathbf{x})]$ in (26), $i = 1, \ldots, k$. Consequently, by Proposition 3.1 and the previous analysis for the case $1 \leqslant r \leqslant r_0$, RLDA is also consistent for all $r_0 < r \leqslant q$.

4.3. Choosing the number of directions in RLDA

By the results of the previous subsection, the reduced linear rule is not consistent for $1 \leqslant r < r_0$, since it ignores directions that are relevant for classification. In contrast, RLDA is consistent when the number of sample canonical coordinates used is $r_0 \leqslant r \leqslant q$. However, since for $r_0 < r \leqslant q$,

$$\|\widehat{\mathbf{y}}_r - \widehat{\mathbf{m}}_{r,i}\|^2 = \|\widehat{\mathbf{y}}_{r_0} - \widehat{\mathbf{m}}_{r_0,i}\|^2 + \widehat{D}_i(r_0, r; \mathbf{x}) + \sum_{j=r_0+1}^{r} [\widehat{\boldsymbol{\gamma}}_j'\widehat{\mathbf{W}}^{-1/2}(\mathbf{x} - \overline{\mathbf{x}})]^2 ,$$

and $\widehat{D}_i(r_0, r; \mathbf{x}) \to 0$, $i = 1,\ldots, k$, the consistency for $r > r_0$ is achieved at the cost of considering *spurious* directions with no effective separatory power.

As mentioned in the introduction, an important issue in RLDA is choosing in practice the number of canonical coordinates. The analysis above suggests that, in general, the choice $r = r_0 = r(\mathbf{B})$ can be recommended. That is, the effective number of sample canonical

coordinates used in (4) should coincide with the exact number of canonical coordinates in the population. However, for problems with $k > 2$ groups, the rank $r_0 = r(\mathbf{B}) \leqslant q = r(\widehat{\mathbf{B}})$ is in general an unknown constant, and its true value should be assessed by some formal testing method. McLachlan [24, Section 6.5.2] reviews inference techniques for $r_0 = r(\mathbf{B})$. Commonly, these testing procedures require $f_i(\mathbf{x}) \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}), i = 1, \ldots, k$. An alternative method is to proceed by trial and error (see e.g. [14, p. 85]) since, as mentioned by Hastie et al. [16, p. 1256], in applications it is often enough to consider a low number $r \leqslant 3$ of sample canonical coordinates, even in problems with a large number of groups.

The trial and error approach is related to selecting the number of canonical coordinates after an exploratory analysis of the relative size of the sum $\sum_{j=1}^{r} \widehat{\lambda}_j$ as compared to the total sum $\sum_{j=1}^{q} \widehat{\lambda}_j$. Specifically, recalling the properties of the spectral decompositions (8) and (24), let $\widehat{p}_j = \widehat{\lambda}_j / \sum_{j=1}^{q} \widehat{\lambda}_j$ be the proportion of total separation among standardized class centroids $\widehat{\mathbf{W}}^{-1/2}(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})$, $i = 1, \ldots, k$, provided by direction $\widehat{\boldsymbol{\gamma}}_j$ and, for $1 \leqslant r \leqslant q$, consider the cumulative proportion $\widehat{q}_r = \sum_{j=1}^{r} \widehat{p}_j = \sum_{j=1}^{r} \widehat{\lambda}_j / \sum_{j=1}^{q} \widehat{\lambda}_j$. A natural criterion for selecting the number of directions as a function of the training data set $\mathbf{D}_n$ is

$$\widehat{r}_0 = \widehat{r}_0(\mathbf{D}_n) = \text{first integer } 1 \leqslant r \leqslant q \text{ such that } \widehat{q}_r \geqslant C\,, \tag{30}$$

where $C > 0$ is a constant close to one. In fact, for an adequate choice of C, the consistency of a *feasible* RLDA rule of the form

$$\|\widehat{\mathbf{y}}_{\widehat{r}_0} - \widehat{\mathbf{m}}_{\widehat{r}_0,i}\|^2 = \min_{1 \leqslant j \leqslant k} \|\widehat{\mathbf{y}}_{\widehat{r}_0} - \widehat{\mathbf{m}}_{\widehat{r}_0,j}\|^2, \tag{31}$$

can be established. To do this, consider the population cumulative separation proportions $q_r = \sum_{j=1}^{r} \lambda_j / \sum_{j=1}^{r_0} \lambda_j$, $r = 1,\ldots, r_0$, and, for convenience, put $q_0 = 0$.

Theorem 4.1. *Under an inverse location model* (18)–(19) *with finite second order moments, the feasible RLDA rule of* (30) *and* (31) *is strongly consistent for all values of C such that* $q_{r_0-1} < C < q_{r_0} = 1$.

Proof. Write $\widehat{l}_n(\mathbf{x}) = \sum_{i=1}^{k} i\, I_{\widehat{R}_{0i,n}}(\mathbf{x})$ for the feasible reduced linear rule, where $\widehat{R}_{0i,n} = \{\mathbf{x} : \|\widehat{\mathbf{y}}_{\widehat{r}_0} - \widehat{\mathbf{m}}_{\widehat{r}_0,i}\|^2 = \min_{1 \leqslant j \leqslant k} \|\widehat{\mathbf{y}}_{\widehat{r}_0} - \widehat{\mathbf{m}}_{\widehat{r}_0,j}\|^2\}$, $i = 1, \ldots, k$, and put $\widehat{l}_0(\mathbf{x})$ for the *theoretical* RLDA rule of (4) based on $r_0 = r(\mathbf{B})$ canonical coordinates. Consider also a sequence $\mathbf{D}_\infty = \{(\mathbf{x}_j, \mathbf{g}_j) : j \geqslant 1\}$ of independent observations with the same distribution than the pair $(\mathbf{x}, \mathbf{g})$. If $(\mathbf{x}, \mathbf{g})$ and $\mathbf{D}_\infty$ are independent, using standard properties of conditional expectation (see e.g. [22, p. 365]), the conditional probability of error $L_n = 1 - P[\widehat{l}_n(\mathbf{x}) = \mathbf{g} \mid \mathbf{D}_n]$ can be represented as

$$L_n = 1 - E[I_{\{0\}}(\widehat{l}_n(\mathbf{x}) - \mathbf{g}) \mid \mathbf{D}_n] = 1 - E[I_{\{0\}}(\widehat{l}_n(\mathbf{x}) - \mathbf{g}) \mid \mathbf{D}_\infty]\,, \tag{32}$$

where $\mathbf{D}_n = \{(\mathbf{x}_j, \mathbf{g}_j) : 1 \leqslant j \leqslant n\}$ is the training data set and $I_{\{0\}}(.)$ is the indicator function of $\{0\} \subset \mathbb{R}$. As in (32), the conditional probability of error of $\widehat{l}_0(\mathbf{x})$ can be written $L_n(r_0) = 1 - E[I_{\{0\}}(\widehat{l}_0(\mathbf{x}) - \mathbf{g}) \mid \mathbf{D}_\infty]$. As seen previously, $L_n(r_0) \to L^*$ a.e., so to get convergence of L_n to L^* it is then enough to proof that, as $n \to \infty$, $L_n - L_n(r_0) \to 0$, a.e.

If $I_{A_n}(.)$ is the indicator function of $A_n = \{\mathbf{D}_n : \widehat{r}_0 = \widehat{r}_0(\mathbf{D}_n) = r_0 = r(\mathbf{B})\}$, the feasible rule $\widehat{l}_n(\mathbf{x})$ can be decomposed additively in the form

$$\begin{aligned}\widehat{l}_n(\mathbf{x}) &= \widehat{l}_n(\mathbf{x}) I_{A_n}(\mathbf{D}_n) + \widehat{l}_n(\mathbf{x}) I_{A_n^c}(\mathbf{D}_n) \\ &= \widehat{l}_0(\mathbf{x}) I_{A_n}(\mathbf{D}_n) + \widehat{l}_n(\mathbf{x}) I_{A_n^c}(\mathbf{D}_n) = \widehat{l}_0(\mathbf{x}) + Z_n \,, \end{aligned} \tag{33}$$

where $Z_n = Z_n(\mathbf{x},\mathbf{D}_n) = [\widehat{l}_n(\mathbf{x}) - \widehat{l}_0(\mathbf{x})] I_{A_n^c}(\mathbf{D}_n)$. By the representations of the conditional probabilities of misclassification L_n and $L_n(r_0)$ given above, (33) implies that the difference $L_n - L_n(r_0)$ can be written in the form

$$L_n - L_n(r_0) = E(W_n \mid \mathbf{D}_\infty) \,, \tag{34}$$

where $W_n = I_{\{0\}}(\widehat{l}_n(\mathbf{x}) - \mathbf{g}) - I_{\{0\}}(\widehat{l}_0(\mathbf{x}) - \mathbf{g}) = I_{\{0\}}([\widehat{l}_0(\mathbf{x}) - \mathbf{g}] + Z_n) - I_{\{0\}}(\widehat{l}_0(\mathbf{x}) - \mathbf{g})$. Observe that $|W_n| \leqslant 1$ so, by (34) and the dominated convergence theorem for conditional expectations (see e.g. [30, p. 216]), to get $L_n - L_n(r_0) \to 0$ a.e. it is then enough to verify that, as $n \to \infty$, $W_n \to 0$, a.e.

From the definition of W_n, for all $\varepsilon > 0$

$$P[\sup_{m \geqslant n} |W_m| \geqslant \varepsilon] \leqslant P[\textstyle\bigcup_{m=n}^{\infty} \{Z_m \neq 0\}] \leqslant P[\bigcup_{m=n}^{\infty} A_m^c] \,, \tag{35}$$

so it suffices to check that the upper bound of (35) converges to zero as $n \to \infty$. By (30) $A_n = \{\mathbf{D}_n : \widehat{r}_0 = \widehat{r}_0(\mathbf{D}_n) = r_0 = r(\mathbf{B})\} = \bigcap_{r=0}^{r_0-1} \{\mathbf{D}_n : \widehat{q}_r < C\} \cap \{\mathbf{D}_n : \widehat{q}_{r_0} \geqslant C\}$, so for any $0 < \alpha < \min\{C - q_{r_0-1}, q_{r_0} - C\} = \min\{C - q_{r_0-1}, 1 - C\}$, the inequality

$$P[\sup_{m \geqslant n} \max_{0 \leqslant r \leqslant r_0} |\widehat{q}_r - q_r| \leqslant \alpha] \leqslant P[\textstyle\bigcap_{m=n}^{\infty} A_m] \tag{36}$$

holds, where $\widehat{q}_0 = q_0 = 0$. Since $\widehat{\lambda}_j \to \lambda_j$ a.e., $j = 1, \ldots, p$, then, for all $0 \leqslant r \leqslant r_0 \leqslant q$, $\widehat{q}_r = \sum_{j=1}^{r} \widehat{\lambda}_j / \sum_{j=1}^{q} \widehat{\lambda}_j \to q_r = \sum_{j=1}^{r} \lambda_j / \sum_{j=1}^{r_0} \lambda_j$ a.e. That is, $\max_{0 \leqslant r \leqslant r_0} |\widehat{q}_r - q_r| \to 0$ a.e. and then both the left- and right-hand sides of inequality (36) converge to 1. Going back to (35), $P[\bigcup_{m=n}^{\infty} A_m^c] = 1 - P[\bigcap_{m=n}^{\infty} A_m] \to 0$. □

As a consequence of the proof above, one has

$$P[\sup_{m \geqslant n} |\widehat{r}_0(\mathbf{D}_m) - r_0| \geqslant \varepsilon] \leqslant P[\textstyle\bigcup_{m=n}^{\infty} A_m^c] \to 0 \,,$$

so $\widehat{r}_0 \to r_0 = r(\mathbf{B})$, a.e. That is, the construction of the feasible rule (31) replaces in the theoretical RLDA rule $\widehat{l}_0(\mathbf{x})$ the unknown quantity r_0 by the strongly consistent estimator $\widehat{r}_0$ of (30). Theorem 4.1 extends then to the unknown $r_0 = r(\mathbf{B})$ case, the consistency of RLDA for r_0 known obtained in the previous subsection. This result justifies also asymptotically the usual exploratory practice of considering a number of directions r such that $\widehat{q}_r \geqslant C > 0$, where $C \cong 1$. In other words, a properly defined separatory criterion of the form (30) leads, as in (31), to an asymptotically optimal allocatory rule.

4.4. Example: the vowel recognition data set

The results obtained in Sections 4.2 and 4.3 can explain some properties of RLDA. Since $r_0 = r(\mathbf{B})$ is the *optimal* dimension for classification, a plot of the conditional probability

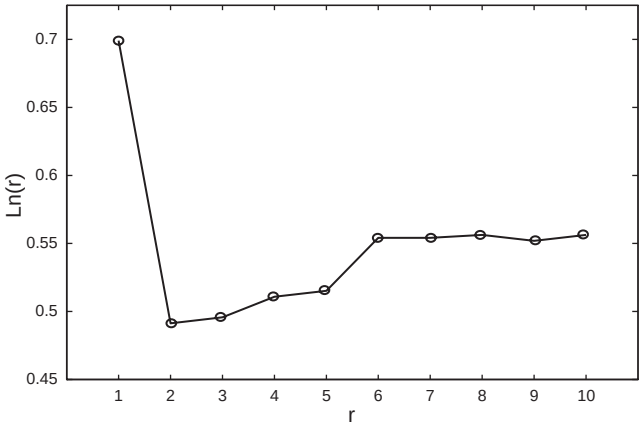

Fig. 1. Vowel recognition data set: estimated error rates $\widehat{L}_n(r)$ in test data for $1 \leqslant r \leqslant 10$.

of error $L_n(r)$ versus r, $r = 1, 2, \ldots, q$, can be conjectured to have a marked decreasing pattern for $1 \leqslant r \leqslant r_0$. After reaching its minimum at $r = r_0$, the plot should have, as a result of the inclusion of spurious directions, an increasing erratic pattern for $r_0 < r$, with a trend to stability as r approaches q.

This is in agreement with the empirical behavior of the plot of the estimated error rates $\widehat{L}_n(r)$ versus r in some well studied classification problems, as for example the *vowel recognition* data set introduced in [16, Section 3]. See also [17, Section 4.3]. For this problem, there are $k = 11$ groups and $p = 10$ variables, and therefore $q = \min(k-1, p) = 10$. There are also $n = 528$ and $m = 462$ observations in, respectively, the training and testing data sets. Fig. 1 displays the estimated error rates in the test data $\widehat{L}_n(r) \cong L_n(r)$, $r = 1, 2, \ldots, 10$. See also Fig. 4.10 in [17, p. 96]. Taking $r_0 = 2$, the plot has a pattern that seems to conform with the characteristics described in the previous paragraph. In particular, $\widehat{L}_n(2) = 0.4913 = \min_{1 \leqslant r \leqslant 10} \widehat{L}_n(r)$ and $\widehat{L}_n(10) = 0.5563$, so RLDA with two canonical coordinates is preferable to LDA.

On the other hand, an analysis of the cumulative proportions $\widehat{q}_r$, $r = 1, 2, \ldots, 10$, leads to $\widehat{q}_1 = 0.5617$, $\widehat{q}_2 = 0.9135$, and $0.9580 \leqslant \widehat{q}_r \leqslant 1$, $3 \leqslant r \leqslant 10$, so the feasible reduced linear rule corresponding to the constant $C = 0.90$ uses $\widehat{r}_0 = 2$ canonical coordinates. This is in agreement with the optimality of the two dimensional RLDA rule observed in the previous paragraph.

5. Final comments

The behavior of LDA and QDA under non-gaussian conditions has been studied extensively, both theoretically and by simulation, in the literature of discriminant analysis. For example, Lachenbruch et al. [21] concentrate on continuous nonnormal data, while Krzanowski [20] analyzes the behavior of LDA under mixtures of binary and continuous data. McLachlan [24, Section 5.6] gives a comprehensive account of additional references. Robustness of LDA and QDA has received recent attention in [5], who study the connec-

tion of LDA and QDA with, respectively, the dimension reduction methods *sliced inverse regression* (SIR) of Li [23] and *sliced average variance estimation* ($SAVE$) of Cook and Weisberg [4]. Hastie and Zhu [18] provide additional insights on the relationships LDA–SIR and QDA–SAVE.

This paper obtains some limit results for the conditional probability of error of the linear and quadratic rules that might be helpful to explain analytically some aspects of their behavior in applications. Our asymptotic theory is based on some adequate pseudo plug-in representations of LDA and QDA, given in Section 2.3, that allow a wider analysis than the one based on standard consistency results for parametric plug-in rules. For continuous data, the results of Section 3 indicate that, when the class prior probabilities are identical and the sample size n is large enough, the correct behavior of LDA and QDA is not related to the usual gaussian assumptions for the class conditional densities $f_i(\mathbf{x})$, but to the more general elliptical families of (18) and (21), respectively. In other words, it is (elliptical) symmetry, and not normality or any other known parametric model, the key aspect for a correct performance in practice of the linear and quadratic rules. As explained above, this can be interpreted as a certain nonparametric character of LDA and QDA.

The results of Section 3 can justify, at least in the common symmetric case, the comments of Hastie et al. [17, p. 89] on the *frequent* correct performance in applications of the linear and quadratic rules. These comments are based on the results reported in the STATLOG project by Michie et al. [25]. In fact, Hastie et al. [17, p. 89] recommend using LDA and QDA as initial simple exploratory classification tools, despite the eventual utilization of more sophisticated classifiers. In practice, the true model for the class conditional densities will not be known. However, according to Section 3, if the available sample information on the $f_i(\mathbf{x})$ through the elements in training sample $\mathbf{D}_n$, $i = 1, \ldots, k$, does not seem to indicate symmetry, the use of the linear and quadratic rules is perhaps questionable.

Resorting to asymptotics is justified by the typical use in practice of moderate to large data sets. However, as pointed out by one of the referees, there is often the situation in which a sample rule must be formed from small data sets, due to the lack of training observations of known classification. If this is the case, our limit results should be then interpreted with some caution.

Finally, the asymptotic results obtained in Section 4.2 detect $r_0 = r(\mathbf{B})$ as an optimal dimension for reduced linear classification. We believe that these results are new and can offer, as related in Section 4.4, some analytical explanation for the behavior in practice of the conditional probability of misclassification of RLDA. In addition to this, Theorem 4.1 can offer, as developed in Section 4.3, some guidelines for choosing the number of directions in RLDA. The idea is to exploit the asymptotic optimality of the separatory criterion (30). Since this method for estimating r_0 is developed under the location model (19), it is perhaps more flexible than methods based on standard tests of dimension that require explicit use of the normality assumption $f_i(\mathbf{x}) \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, \ldots, k$.

Acknowlegdments

The authors are grateful to the Editor and two anonymous referees for their careful review of the first version of this paper.

References

[1] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, second ed., Wiley, New York, 1984.
[2] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth, Bellmont, CA, 1984.
[3] N.A. Campbell, Canonical variate analysis, Austral. J. Statist. 26 (1) (1984) 86–96.
[4] R.D. Cook, S. Weisberg, Discussion of Sliced Inverse Regression for Dimension Reduction by Li (1991), J. Amer. Statist. Assoc. 86 (1991) 328–332.
[5] R.D. Cook, X. Yin, Dimension reduction and visualization in discriminant analysis (with discussion), Austral. N. Zealand J. Statist. 43 (2) (2001) 147–199.
[6] P.W. Cooper, Statistical classification with quadratic forms, Biometrika 50 (1963) 439–448.
[7] N.E. Day, Linear and quadratic discrimination in pattern recognition, IEEE Trans. Inform. Theory IT-15 (1976) 419–420 (1969).
[8] L. Devroye, L. Györfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer, New York, 1996.
[9] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugenics 7 (1936) 179–188.
[10] B. Flury, A First Course in Multivariate Analysis, Wiley, New York, 1997.
[11] N. Glick, Sample-based classification procedures derived from density estimators, J. Amer. Statist. Assoc. 67 (1972) 116–122.
[12] N. Glick, Consistency conditions for probability estimators and integrals of density estimators, Utilitas Math. 6 (1974) 61–74.
[13] N. Glick, Sample-based classification procedures related to empiric distributions, IEEE Trans. Inform. Theory IT-22 (1976) 454–461.
[14] R. Gnanadesikan, Methods for Statistical Data Analysis of Multivariate Observations, second ed., Wiley, New York, 1997.
[15] T. Hastie, R. Tibshirani, Discriminant analysis by Gaussian mixtures, J. Roy. Statist. Soc. Ser. B 58 (1996) 155–176.
[16] T. Hastie, R. Tibshirani, A. Buja, Flexible discriminant analysis by optimal scoring, J. Amer. Statist. Assoc. 89 (1994) 1255–1270.
[17] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer, New York, 2001.
[18] T. Hastie, M. Zhu, Discussion of Dimension Reduction and Visualization in Discriminant Analysis, by Cook and Yin, Austral. N. Zealand J. Statist. 43 (2) (2001) 179–185.
[19] R.A. Johnson, D.W. Wichern, Applied statistical Multivariate Analysis, fifth ed., Prentice-Hall, Upper Saddle River, NJ, 2002.
[20] W.J. Krzanowski, The performance of Fisher's linear discriminant function under non-optimal conditions, Technometrics 19 (1977) 191–200.
[21] P.A. Lachenbruch, C. Sneeringer, L.T. Revo, Robustness of the linear and quadratic discriminant functions to certain types of non-normality, Comm. Statist. 1 (1973) 39–57.
[22] R.G. Laha, V.K. Rohatgi, Probability Theory, Wiley, New York, 1979.
[23] K.C. Li, Sliced inverse regression for dimension reduction (with discussion), J. Amer. Statist. Assoc. 86 (1991) 316–342.
[24] G.J. McLachlan, Discriminant Analysis and Statistical Pattern Recognition, Wiley, New York, 1992.
[25] D. Michie, D. Spiegelhalter, C. Taylor (Eds.), Machine Learning, Neural and Statistical Classification, Ellis Horwood Series in Artificial Intelligence, Ellis Horwood, Chichester, UK, 1994.
[26] R.J. Muirhead, Aspects of Multivariate Statistical Theory, Wiley, New York, 1982.
[27] B.L.S. Prakasa Rao, Nonparametric Functional Estimation, Academic Press, New York, 1983.
[28] C.R. Rao, Advanced Statistical Methods in Biometric Research, Wiley, New York, 1952.
[29] G.A.F. Seber, Multivariate Observations, Wiley, New York, 1984.
[30] A.N. Shiryayev, Probability, Springer, New York, 1984.
[31] D.E. Tyler, Asymptotic Inference for Eigenvectors, Ann. Statist. 9 (1981) 725–736.